

Validace shlukování

18. dubna 2016

Proč validovat shluky:

- detekce šumu,
- porovnání shlukovacích algoritmů,
- porovnání dvou množin shluků,
- porovnání dvou shluků.

Numerické míry (indexy) používané k posouzení různých hledisek platnosti shluků se dělí do těchto tří typů:

Externí indexy měří v jakém rozsahu se shluky shodují s externí informací (rozhladem, oštitkováním).

Interní indexy měří kvalitu struktury shluků bez externí informace.

Relativní indexy porovnávají dvě různá shlukování nebo shluky.

1 Externí indexy

Mějme shlukování $\mathcal{C}$ a externí informaci (rozklad) $\mathcal{P}$:

$$\begin{aligned}\mathcal{C} &= \{C_i \mid 1 \leq i \leq c\} \\ \mathcal{P} &= \{P_j \mid 1 \leq j \leq d\}\end{aligned}$$

A označme: n – počet objektů, $n_i = |C_i|$, $n_j = |P_j|$, $n_{i,j} = |C_i \cap P_j|$, $p_i = \frac{n_i}{n}$, $p_j = \frac{n_j}{n}$, $p_{i,j} = \frac{n_{i,j}}{n}$.

1.1 Randův index (Rand index)

Randův index je definován jako

$$R = \frac{\binom{n}{2} - \sum_i \binom{n_i}{2} - \sum_j \binom{n_j}{2} + 2 \cdot \sum_{i,j} \binom{n_{i,j}}{2}}{\binom{n}{2}}.$$

Označme

- a – počet dvojic (x, x') různých objektů, t.ž. x, x' jsou ve stejném shluku v $\mathcal{C}$ i ve stejné skupině v $\mathcal{P}$.
- b – počet dvojic (x, x') různých objektů, t.ž. x, x' jsou ve stejném shluku v $\mathcal{C}$, ale v různých skupinách v $\mathcal{P}$.
- c – počet dvojic (x, x') různých objektů, t.ž. x, x' jsou v různých shlucích v $\mathcal{C}$, ale ve stejné skupině v $\mathcal{P}$.
- d – počet dvojic (x, x') různých objektů, t.ž. jsou v různých shlucích v $\mathcal{C}$ i v různých skupinách v $\mathcal{P}$.

Tedy

$$\begin{aligned}
 a &= \sum_{i,j} \binom{n_{ij}}{2} \\
 b &= \sum_i \binom{n_i}{2} - \sum_{i,j} \binom{n_{ij}}{2} = \sum_i \binom{n_i}{2} - a, \\
 c &= \sum_j \binom{n_j}{2} - \sum_{i,j} \binom{n_{ij}}{2} = \sum_j \binom{n_j}{2} - a, \\
 d &= \binom{n}{2} - \sum_i \binom{n_i}{2} - \sum_j \binom{n_j}{2} + \sum_{i,j} \binom{n_{ij}}{2} = \binom{n}{2} - a - b - c.
 \end{aligned}$$

Randův index můžeme vyjádřit jako

$$R = \frac{a + d}{a + b + c + d}.$$

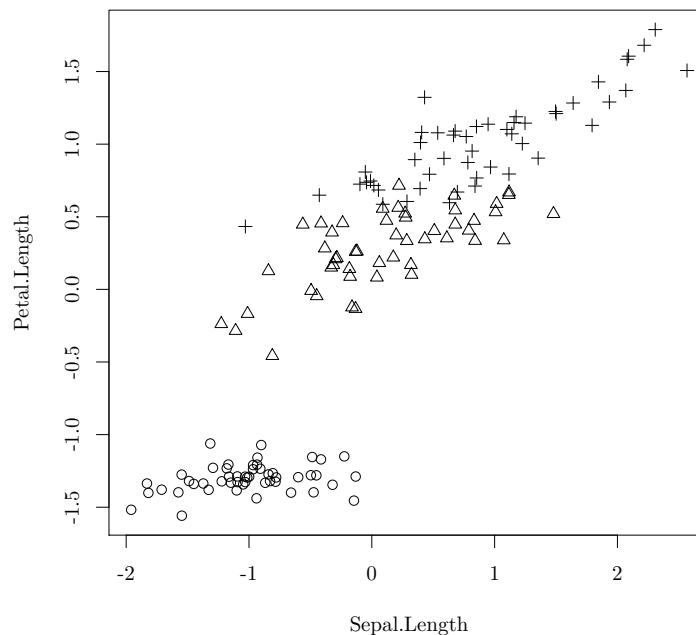
Randův index nabírá hodnot z intervalu $[0, 1]$, vyšší index představuje vyšší shodu.

```

RandIndex <- function(C,P)
{
  M = choose(length(C),2);
  return(
    (M - sum(choose(table(C),2)) - sum(choose(table(P),2)) + 2 * sum(choose(table(C,P),2))) / M)
  )
}

> data <- apply(iris[,c(1,3)],c(1,2),FUN=jitter)
> data <- scale(data)
> P <- c(rep(1,50),rep(2,50),rep(3,50))
> plot(data,pch=P)

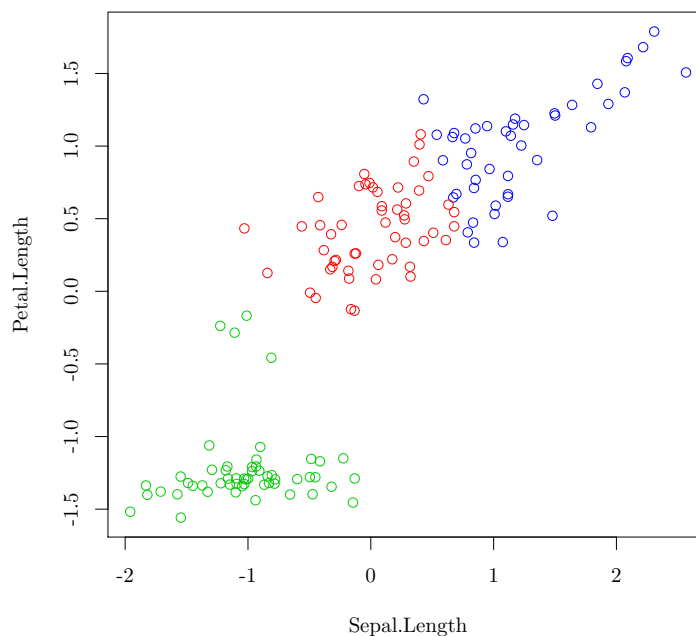
```



```

> C <- kmeans(data,3)
> plot(data,col=C)

```



```
> RandIndex(C,P)
[1] 0.802774
```

1.2 Jaccardův index (Jaccard index)

Jaccardův index je definován takto

$$J = \frac{\sum_{i,j} \binom{n_{ij}}{2}}{\sum_i \binom{n_i}{2} + \sum_j \binom{n_j}{2} - \sum_{i,j} \binom{n_{ij}}{2}}.$$

S použitím počtů a, b, c, d z předchozí části můžeme Jaccardův index vyjádřit jako

$$J = \frac{a}{a + b + c}.$$

Výpočet Jaccardova indexu:

```
JaccardIndex <- function(C,P)
{
  A2 = sum(choose(table(C,P),2));

  return(A2/(
    sum(choose(table(C),2))
    +
    sum(choose(table(P),2))
    - A2))
}
```

Aplikace na předchozí příklad:

```
> JaccardIndex(C,P)
[1] 0.5399708
```

1.3 F -míra (harmonický průměr; F -measure)

Přesnost a úplnost:

$$\text{Precision}(i, j) = \frac{n_{i,j}}{n_j} \quad \text{Recall}(i, j) = \frac{n_{i,j}}{n_i}$$

$$F_1 = \frac{2 \cdot \text{Recall}(i, j) \cdot \text{Precision}(i, j)}{\text{Precision}(i, j) + \text{Recall}(i, j)}$$

Hodnoty F -measure jsou v intervalu $[0,1]$ a větší hodnoty indikují vyšší kvalitu shlukování.

```
precision <- function(C,P,i,j)
{
  sum(P==j & C==i)/sum(P==j)
}

recall <- function(C,P,i,j)
{
  sum(P==j & C==i)/sum(C==i)
}

F1 <- function(i,j,C,P)
{
  rec <- recall(i,j,C,P);
  pre <- precision(i,j,C,P);
  if ((rec + pre)>0)
  {
    2 * pre * rec / (rec + pre)
  } else 0
}

> F1(1,2,C,P)
[1] 0.72

> F1(1,3,C,P)
[1] 0.32

> F1(2,3,C,P)
[1] 0
```

Zobecnění umožňující preferovat přesnost nebo úplnost.

$$F_\beta = (1 + \beta^2) \cdot \frac{\beta^2 \cdot \text{Recall}(i, j) \cdot \text{Precision}(i, j)}{\text{Precision}(i, j) + \text{Recall}(i, j)}$$

1.4 Fowlkes & Mallows index

Fowlkes & Mallows index je definován jako

$$FM = \frac{\sum_{i,j} n_{ij}^n - n}{\sqrt{(\sum_i n_i^2 - n)(\sum_j n_j^2 - n)}}.$$

S použitím počtů a, b, c, d můžeme Fowlkes & Mallows index vyjádřit takto:

$$FM = \sqrt{\frac{a}{a+b} \cdot \frac{a}{a+c}} = \frac{a}{\sqrt{(a+b) \cdot (a+c)}}.$$

Nabývá hodnot z intervalu $[0, 1]$.

```
FM <- function(C,P)
{
  A = sum(choose(table(C,P),2));
  AaB = sum(choose(table(C),2));
  AaC = sum(choose(table(P),2));

  return(sqrt(A/(AaB * AaC)))
}
```

Aplikace na předchozí příklad.

```
> FM(C,P)
[1] 0.7012791
```

Poznámka: Fowlkes a Mallows uvedli toto kritérium pro porovnávání hierarchických shlukování.

1.5 Entropie (entropy)

Entropie shluku j -tého shluku.

$$E_j = - \sum_{i=1}^c \frac{n_{ij}}{n_j} \cdot \log_2 \frac{n_{ij}}{n_j},$$

Celková entropie

$$E = \sum_{j=1}^d \frac{n_j}{n} E_j.$$

Nabývá hodnot z intervalu $[0, \log_2 c]$.

1.6 Čistota (purity)

Použijeme terminologii zavedenou u entropie. Čistota j -tého shluku je daná

$$P_j = \max_{i=1}^c p_{ij}$$

a celková čistota shlukování

$$P = \sum_{j=1}^c p_j P_j.$$

Nabývá hodnot z intervalu $(0, 1]$.

1.7 Vzájemná informace (mutual information)

$$MI = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{p_i p_j}$$

Nabývá hodnot z intervalu $[0, \log_2 c]$.

2 Interní indexy

2.1 Calinského-Harabaszův validační index (Calinski-Harabasz validity index)

Calinského-Harabaszův validační index je definován jako

$$CH = \frac{\text{trace}(S_B)}{\text{trace}(S_W)} \cdot \frac{N - c}{c - 1},$$

kde S_W je matice rozptylu uvnitř shluků a S_B je matice rozptylu mezi shluky.

$$S_W = \sum_{i=1}^c S_i, \text{ kde } S_i = \sum_{\vec{x} \in C_i} (\vec{x} - \vec{\mu}_i)(\vec{x} - \vec{\mu}_i)^T$$

$$S_T = \sum_{\vec{x} \in C_i} (\vec{x} - \vec{\mu})(\vec{x} - \vec{\mu})^T$$

Matici celkového rozptylu můžeme rozložit na součet matice rozptylu mezi shluky a matice rozptylu uvnitř shluků

$$S_T = S_B + S_W.$$

S_B můžeme odhadnout jako

$$S_B = \sum_{i=1}^c n_i (\vec{\mu}_i - \vec{\mu})(\vec{\mu}_i - \vec{\mu})^T.$$

2.2 Daviesův-Bouldinův validační index (Davies-Bouldin validity index)

$$BD = \frac{1}{c} \sum_{i=1}^c \max_{i \neq j} \left\{ \frac{d(C_i) + d(C_j)}{d(C_i, C_j)} \right\}$$

kde $d(C_i)$ průměrná vzdálenost objektů ve shluku od středu shluku. $d(C_i, C_j)$ vzdálenost mezi středy shluků.

2.3 Dunnův validační index (Dunn index)

$$D = \min_{1 \leq i \leq c} \left\{ \min_{1 \leq i \leq c} \left\{ \frac{d(C_i, C_j)}{\max_{1 \leq k \leq c; i \neq j} d(C_k)} \right\} \right\}$$